C-MIST: An Automated Oceanographic Data Processing Software Suite

A. Pruessner
National Oceanic and Atmospheric Administration/ National Ocean Service / CO-OPS, 1305 East West Hwy, Silver Spring, MD 20910,
Armin.Pruessner@noaa.gov

P. Fanelli
National Oceanic and Atmospheric Administration/ National Ocean Service / CO-OPS, 1305 East West Hwy, Silver Spring, MD 20910,
Paul.Fanelli@noaa.gov

C. Paternostro
National Oceanic and Atmospheric Administration/ National Ocean Service / CO-OPS, 1305 East West Hwy, Silver Spring, MD 20910,
Christopher.Paernostro@noaa.gov

Abstract - **We introduce an enterprise software suite to automate the processing of large-scale, large-volume current meter data for dissemination to the nation's maritime community. Currents Measurement Interface for the Study of Tides (C-MIST) is a web-based end-to-end state-of-the-art data management system to ingest, quality control, analyze, and disseminate water velocity and related data from coastal and estuarine collections. Using real-world case studies from Southeast Alaska and Galveston Bay we demonstrate how the various modules of the system allow oceanographers to apply their expertise in analyzing complex data to research natural phenomena.**

I. INTRODUCTION

Currents Measurement Interface for the Study of Tides (C-MIST, online at cmist.noaa.gov) is a web-based end-to-end state-of-the-art data management system to ingest, quality control, analyze, and disseminate water velocity and related data. It has been operational since January 2007 for processing of NOAA current meter data.

The motivation for the development of the system involved several factors, including the extremely large amounts of data that need to be processed yearly after each field season. A single survey data set of one month may contain as many as 250,000 records of data, resulting in several million measured values. A field season may include 75 current meter deployments (with the number projected to grow to 140 in the coming years). Other factors for the development of C-MIST include inconsistency of results obtained by older data processing methods, inflexibility of some of the older modules, difficulty in reproducing some of the results, and complexity in extending available analysis types. Furthermore, the existing methods did not allow for much research of the data beyond the standard operating procedures of analysis.

The benefits of the C-MIST software suite include streamlining preliminary analysis, thus allowing time and resources for more in-depth investigations of the physical phenomena, increasing consistency of results between users, reproducibility of results, and improving overall data quality.

In this paper we discuss current meter data processing at NOAA's National Ocean Service (NOS), give an overview of the main modules of the software, describe the system architecture, including the software models used to make the project a success, and show analysis case studies of recent real-world sensor deployment in Alaska and the Galveston Bay areas. Although we touch upon software quality issues within C-MIST, these are described in more detail in [2] and [3].

II. OCEANOGRAPHIC DATA ANALYSIS AT NOAA

The Center for Operational Oceanographic Products and Services (CO-OPS) at NOS has the mandate to provide the nation with accurate observed water-level information and tide and tidal current predictions for the U.S. coasts. CO-OPS installs and maintains instruments that take accurate measurements of water level, current, meteorological and air gap data. The incoming water level and current data are quality controlled and analyzed to provide curve fit coefficients used to create predictions. Traditional tidal harmonic analyses of the observations are used to compute the coefficients in response to forcing of the earth-moon-sun gravitational system. These coefficients are then used to create tidal predictions [4].

Prior to the development of C-MIST, data processing at CO-OPS was done either manually by oceanographers on a workstation interacting with database queries, file transfers, running analysis routines, and other individual applications or semiautomatically for certain web-based product applications. Depending on the application, the processes involve sequential use of separate programs and procedures and, often, separate machines. Some of the more intensive manual procedures require oceanographer expertise to analyze and accept results (for instance, the procedure for analysis of the seasonal constituents and the process for rejection of certain low amplitude constants for use in the tide prediction equation). See [6] for more details.

The multitude of steps, as well as the different processing hardware and environments, were both prone to human error and time-consuming. Indeed, to process a typical station (one month deployment and 25 bins or

depth measurements) prior to C-MIST may have required up to 8 hours of processing time (plus 2 days to FedEx the raw data disk to NOS). To process an entire field season consisting of 75 data sets would require roughly 75 working days for a single data processor. With future projections of up to 140 station deployments annually, the need for a more efficient data processing method became more urgent.

III. SOFTWARE DESIGN, PROCESSES, AND MODULES

The C-MIST software suite follows a modular approach and uses well accepted software quality models. This includes formal processes for requirements elicitation, system design, development, testing and training (see the software lifecycle in Figure 1). The quality models used are not discussed in detail, however it is shown that the structured approach and modularity of components allowed for high quality software to be developed on schedule and with desired functionality. Furthermore, the lifecycle follows a phased approach with multiple releases incorporating new features and bug fixes at each step. For more information on the software quality processes used within this project, see [2] and [3].

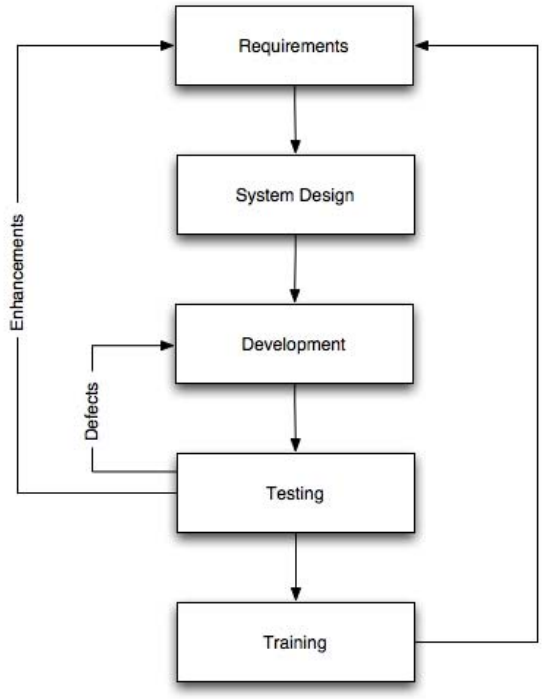

Figure 1. C-MIST software lifecycle.

C-MIST consists of various modules, including

- data ingestion module
- quality assurance (QA) module
- data analysis module
- data visualization module and
- report generation module

The modules interact with the main graphical user interface (GUI) using uniform Application Programming Interfaces (API). APIs allow for consistent communication and allow new components to be added easily. For example, the addition of new plots or new analyses generally does not require much overhead in terms of software development since communication for each type is similar. We describe some of the main modules of the software below.

A. Raw Data Quality Control and Data Ingestion

The data usually comes in a binary, vendor-specific format and needs to be converted into columnar ASCII format. Although all raw data will be ingested into the system database, data committed must first be quality controlled. This includes checking for missing data records, extreme instrument pitch and roll levels, determining valid instrument heading, water velocity, and signal strength for all measurements. All data are ingested into the database with data not satisfying flagged appropriately. Furthermore, during ingestion data may be converted as needed for uniformity to standard engineering units.

B. Data Analysis

The general notion of the data analysis phase is to help oceanographers analyze large data sets and be able to extrapolate future tidal movements based on the available data. Arguably, the most common technique for processing of sequential data is time-series analysis. The idea of harmonic analysis is to define the periodic time series in terms of a finite number of dominant period functions [1], [4]. CO-OPS makes use of various methods for fitting a finite sequence of cosine terms (i.e. determining amplitude and phase of each term) to the observed data, depending on the amount of data available.

These include:

- Least squares harmonic analysis for more than 180 days of continuous or discontinuous data
- 29 day Fourier harmonic analysis for time series between 29 and 180 days of continuous data, and
- 15 day Fourier harmonic analysis for 15-29 days of continuous data

The resulting constituents (amplitude and phase coefficients) allow for modeling to predict tidal forcing at the station fo rany time. In particular, using the constituent information obtained, CO-OPS makes tidal current predictions for a whole year. These are published in the yearly Tidal Current Tables. See, for example, [5].

C. Data Visualization and Report Generation

C-MIST allows the user to generate a wide variety of plots, such as time series plots (observed, predicted and residual tidal current speeds over a given time period), vector plots (current directions), as well as animations showing information over a wider time period. We also have incorporated mapping capabilities using Google Maps [7] to visually show locations of deployed current meters.

In addition, users have access to standard as well as customized reports generated on-the-fly by the system. The reports include metadata, quality control plots, and analysis results generated during a user session and provided in portable document format (PDF).

IV. CASE STUDIES

This section discusses two different case studies. The first demonstrates the daily use of C-MIST through the study of a deployment in Southeast Alaska with well-defined flood and ebb flow. The second involves an instrument replacement in Galveston Bay near Houston, Texas in which the C-MIST system was used to provide accurate quality-control information in order to determine suitability of data dissemination.

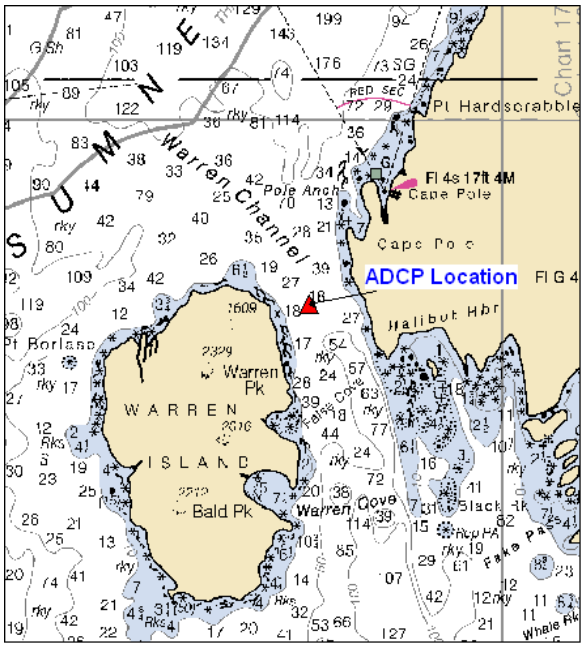

Figure 2. Location of the current meter deployed in Warren Channel during the summer of 2006. Current is forced between Kosciusko Island and smaller Warren Island to the southwest.

A. *Data Collection (Southeast Alaska)*

In the summer of 2006, NOAA/CO-OPS deployed 42 current meters in Southeastern Alaska. Data used for this case study were collected in Warren Channel just off of Sumner Strait. Figure 2 shows the location of this deployment in a channel which lies between Kosciusko Island and smaller Warren Island to the southwest. The data is expected to be rectilinear with well defined flood and ebb directions because flow is constricted between two islands.

A Teledyne RDI Workhorse 300 kHz Acoustic Doppler Current Profiler (ADCP) was used for this study. The ADCP was deployed on the bottom looking upward in about 40 meters of water from for the period of June 15th to August 3rd, 2006. Data were sampled at 3 meter increments from 27.9m depth to the near-surface layer. Following retrieval of the instrument, data were ingested into the National PORTS Database (NPDB).

B. *Quality Control (Southeast Alaska)*

A C-MIST session begins with station selection and then data retrieval from the NPDB database. Prior to prediction analyses being performed, raw data is first quality controlled. This step is necessary to assure the high quality of input data for analysis.

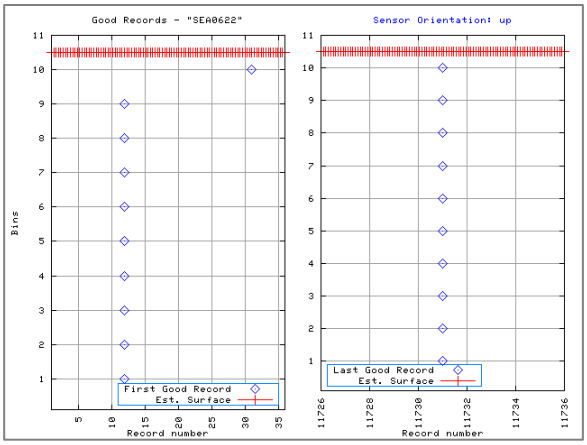

Figure 3. Data QC plot showing where good data begins and ends at each depth. Based on this plot, users can quality-control the data by trimming bad data. Note that bin 10 good data begins around record 31.

This system is uniquely designed for current meter data collected by CO-OPS. This method of current meter deployment produces bad data at each end of the collected time series as the meter is usually turned on prior to stabilizing on the ocean floor and turned off after recovery on deck of the vessl. C-MIST automatically determines the first and last good data record at each depth based on instrument-specific quality control parameters such as echo intensity and and depth (or pressure) values. Figure 3 shows a typical spatial and temporal data quality analysis of an extracted data set. Estimated first and last good records are offerred to the user for trimming.

In addition to removing bad data, the quality control module automatically performs a gap analysis on data at each depth and calculates the last good bin of data. Finally, the quality control module calculates the station depth using pressure data relative to *mean-lower-low water* (mean of the lowest of the low waters for each day due to the declinational effects of the sun and moon).

While C-MIST automates many procedures that users may go through to analyze tidal current data, it is important for oceanographers to be able to visualize the data to make their own quality control assessments. The data visualization module allows users to generate a suite of plots including time-series and scatter plots of many raw data parameters and contour plots to visualize parameters over depth. Manual plots can be produced allowing the user to plot any parameter against any five other parameters of data.

Figure 4 shows a time series of speed and direction at a depth of 6.9 meters for the duration of deployment. The strongest tidal current speeds are often during spring tide. At Warren Channel, these speeds reach upwards of 180 cm/s (~3.5 kt). Examination of the direction time series shows that the current is predominantly flowing in two opposite directions (often referred to as a reversing current). This is expected as the data was located in a clearly defined channel where flood currents are forced up the channel and ebb currents are forced down the channel.

Figure 5 shows a scatter plot of the north and east components of velocity for the same depth. Here we see the data matches up well with the channel orientation seen in Figure 2.

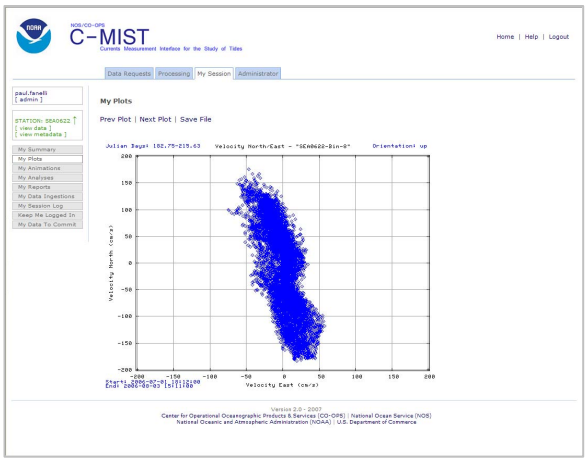

Figure 4. C-MIST Quality Control Plot: time series of speed and direction.

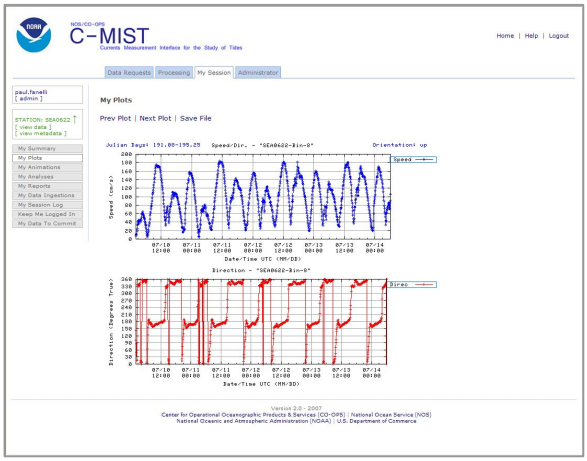

Figure 5. C-MIST Quality Control Plot: time series of speed and direction.

C. Data Analysis and Report Generation (Southeast Alaska)

The tidal current data is now ready to be analyzed using one of the harmonic analysis routines described in Section III. While the Warren Channel data does not contain any gaps, a least squares analysis [6] was determined to be optimal to resolve over-tides. By default, C-MIST automatically runs a suite of routines that CO-OPS uses to analyze tidal current data. The automation uses defaults and best guess parameters to generate quick results for the user by calculating *Greenwich Intervals* (the average time of the max flood and the max ebb relative to the passage of the moon over the Greenwich meridian), determining a harmonic analysis routine to extract the individual tidal constiuent phases and amplitudes, creating a time series prediction and finally plotting overlays of the predicted time series over the observations. C-MIST then allows users to run additional anlyses with user-specified parameters or generating custom plots. Users assess the quality of the predictions and examine the residual to determine if the results are suitable for dissemination to the maritime community. If the default analysis results are not suitable, for example if the residual shows additional harmonics, users have the ability to discard results and rerunning the harmonic analysis with user-defined parameters.

The results of a data analsysis session are summarized in a cumulative report. The system automatically generates a PDF document containing all the analyses results and plots generated, including quality control and prediction plots.

Figure 6 shows a prediction vs observed time series and a scatter plot from the cumulative data anslysis PDF report. The results indicate that data from Warren Channel data matches the predictions pretty closely, however some periodicity in the residual indicates that all of the tidal influence may not be accounted for in the constituents resulting from harmonic analysis. Updated harmonic analysis programs being implemented at NOAA will allow us to analyze for more constituents that will better capture non-linear and frictional effects and should reduce these periodic residuals.

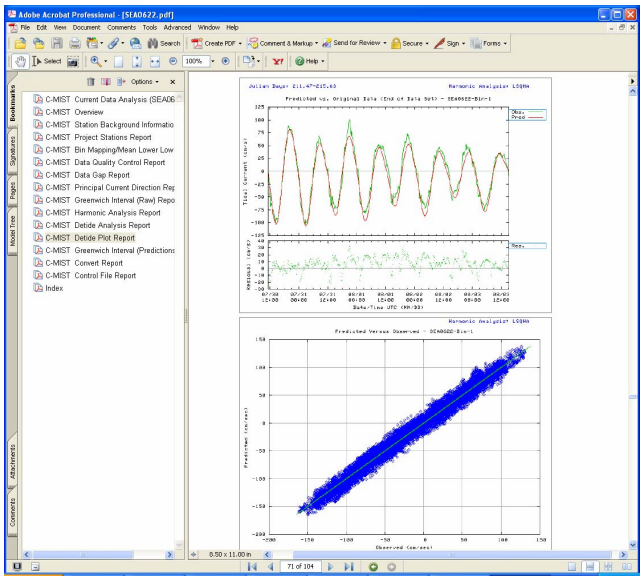

Figure 6. C-MIST PDF report. The top plot shows the relationship between predicted data (red-solid) and the observed data (green-dashed) using the harmonic analysis results and the observed data for a 4 day period during the deployment. The bottom plot shows a perdiceted vs observed scatter plot indicating a strong correlation (fitness).

D. Quality-Control (Galveston Bay)

While C-MIST can be used to perform routine data analysis, it can also be used to diagnose issues with instrumentation for real-time systems. Figure 7 shows a contour plot of echo intensity for a recently deployed 3-beam downward looking current profiler in Galveston Bay along a buoy in the Houston-Galveston Ship Channel. Echo intensity represents the strength of the returning acoustic signal which logically decreases with distance from the source until reaching the seafloor. All three beams show a sharp increase around bin 13 (corresponding to a depth of roughly 15.9m) indicating the expected reflection (red contours) off the hard surface of the seafloor. However in beam 1, there is an unusually strong echo

moving up and down over time at mid-depths. The movement of the buoy changes with the fluctuating current altering the aim of the current meter's beams. In this case, at high current speeds in both the flood and ebb directions, beam 1 produces strong echoes at mid-depths similar to reflections off a hard object. This periodic obstruction in beam 1 correlates with slow speeds at maximum flows. This analysis helped prevent erroneous information from being disseminated to ships within the Houston Ship Channel and guided field crews to narrow the possibilities of problems when servicing this instrument.

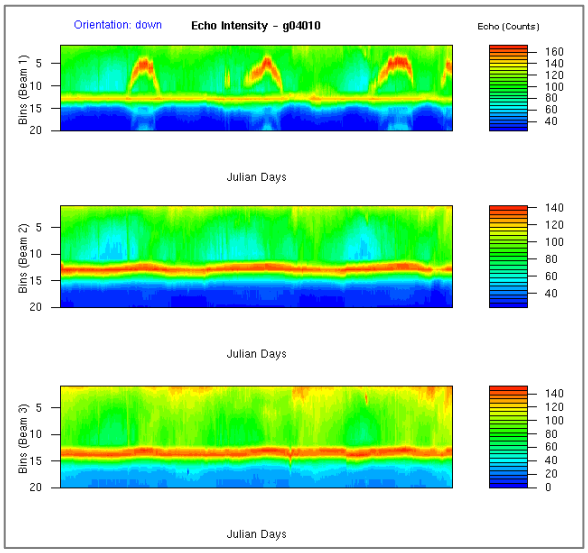

Figure 7. QC echo intensity contour plot for station g04010 (Houston Ship CH Entrance LBB 18) indicating unusual echo intensity spikes for beam 1 (first plot above) compared to the other two beams.

IV. CONCLUSIONS

C-MIST is a state-of-the-art end-to-end oceanographic data analysis system, which improves on present state-of-the-art. The system is designed to be able to handle large-scale data sets that have high spatial and temporal resolution. The data management system provides oceanographers the means to study water velocity data using a wide suite of mathematical and graphical tools allowing them create more products rather than spending time altering data sets and creating analysis program input files. Furthermore, the data processing efficiency has improved by decreased processing time from hours to minutes with respect to previous manual methods.

Future work may include the ability to allow the public to upload their own data sets for processing, the ability to compare results from various analyses, as well as expanded public data access capabilities.

Acknowledgments

The authors would like to thank Steve Gill for insightful comments that improved this paper.

REFERENCES

[1] W.J. Emery, R. E. Thompson, *Data Analysis Methods in Physical Oceanography*, Second and Revised Edition, Elsevier BV, Amsterdam, 2004.

[2] C. Paternostro, A. Pruessner, and R. Semkiw, "Desining a Quality Oceanographic Data Processing Environment", *Proceedings of the MTS/IEEE Oceans Conference*, 3, pp. 2527-2531, Spetember 19-23, Washington DC, 2005.

[3] A. Pruessner and C. Paternostro, Software Quality : From Legacy Codes to Web-based Enterprise Applications, *Software Quality Professional*, 8 (3), 2006.

[4] P. Shureman, *Manual of Harmonic Analysis and Prediction of Tides*, Special Publication No. 98, Revised Edition, US Department of Commerce, Coast and Geodetic Survey, 1940.

[5] Tide Current Tables 2007, Atlantic Coast of North America. US Department of Commerce, National Oceanic and Atmospheric Administration, National Ocean Service.

[6] C. Zervas, Tidal Current Analysis Procedures and Associated Computer Programs,*NOAA Technical Report NOS CO-OPS 0021,* US Department of Commerce, Silver Spring, MD, 1999.

[7] Google Maps API, online at http://www.google.com/apis/maps/documentation/, accessed Aug. 9 2007.